**UNIVERSIDAD DE BUENOS AIRES - 2025**

**BIG DATA & MACHINE LEARNING**

**Trabajo Práctico N° 4**

**Métodos Supervisados: Regresión &**

**Clasificación usando la EHP**

1. *Enfoque de validación*



Para que sea más ilustrativo, muestro el promedio de las 172 variables.

En general, los valores p son mayores a 0.05 en la mayoría de las variables, lo cual sugiere que no hay diferencias estadísticamente significativas entre los conjuntos de entrenamiento y prueba, lo que indica una correcta partición de los datos sin sesgo evidente. Esto es crucial para la validez de los modelos predictivos, ya que garantiza que los modelos estén siendo entrenados y evaluados sobre muestras comparables.

Sin embargo, hay algunas excepciones donde los p-value son bajos (por ejemplo, V3\_M o ADECOCUR), lo que podría indicar diferencias relevantes entre los subconjuntos para esas variables particulares. Estas discrepancias podrían deberse a variabilidad estructural en los datos.

1. *Metodo Supervisado 1: Modelo de Regresión Lineal*

**Tabla 2. Estimación por regresión lineal de salarios usando la base de entrenamiento**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Var. Dep: *salario\_semanal*  Variables | Modelo 1  (1) | Modelo 2  (2) | Modelo 3  (3) | Modelo 4  (4) | Modelo 5  (5) |
| *edad* | 36,207\*\*\* (1,70) | 329,414\*\*\* (5,73) | 233,718\*\*\* (7,04) | 226,208\*\*\* (6,92) | 83,956\*\*\* (10,10) |
| *edad2* | - | -3,702\*\*\* (0,07) | -2,662\*\*\* (0,08) | -2,544\*\*\* (0,08) | -0,946\*\*\* (0,12) |
| *educ* | - | - | 190,809\*\*\* (8,34) | 203,045\*\*\* (8,21) | 187,209\*\*\* (7,58) |
| *Mujer* | - | - | - | -1.888,370\*\*\* (69,72) | -870,564\*\*\* (65,40) |
| *horastrab* | - | - | - | - | 108,500\*\*\* (1,67) |
| PET | - | - | - | - | -638,039\*\*\* (128,39) |
|  |  |  |  |  |  |
| N (observaciones) | 20442 | 20442 | 20442 | 20442 | 20442 |
| *R2* | 0,022 | 0,141 | 0,162 | 0,191 | 0,329 |

Nota: destaque con \*, \*\*, y \*\*\* cuando el p-valor de los coeficientes reportados sean menor que 0.1, 0.05 y 0.001 respectivamente.

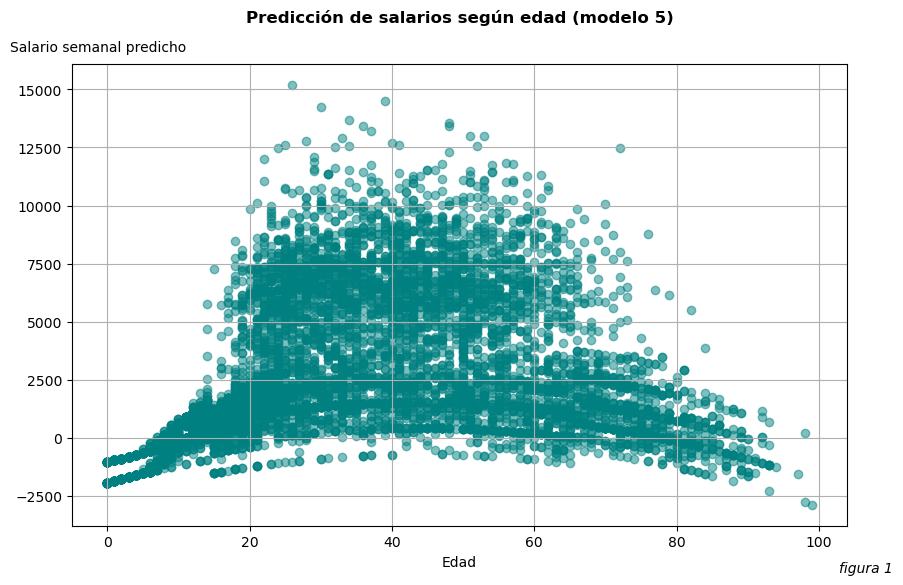
**Tabla 3. Performance por regresión lineal de la predicción de salarios usando la base de testeo**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Var. Dep: *salario\_semanal* | Modelo 1  (1) | Modelo 2  (2) | Modelo 3  (3) | | Modelo 4  (4) | Modelo 5  (5) |
| *MSE test* | 28.257.253,49 | 24.726.123,78 | | 23.940.168,03 | 23.098.062,38 | 19.238.603,89 |
| *RMSE test* | 5.315,75 | 4.972,54 | | 4.892,87 | 4.806,04 | 4.386,18 |
| *MAE test* | 3.421,92 | 3.086,89 | | 3.087,82 | 3.036,43 | 2.453,05 |

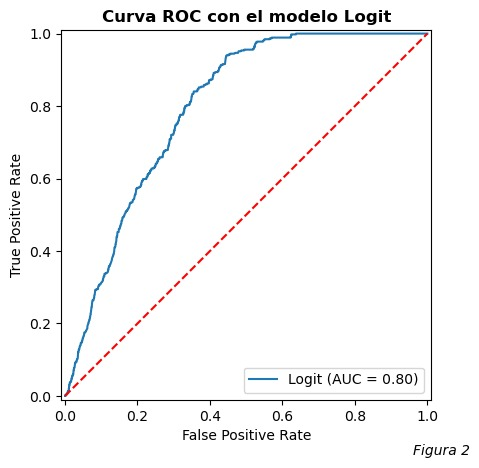
Del conjunto de los modelos, el 5 es el más completo y el que logra explicar aproximadamente el 33% de la variabilidad en el salario semanal. Este modelo incluye variables como edad, edad al cuadrado, educación, género, horas trabajadas y participación en el programa PET.

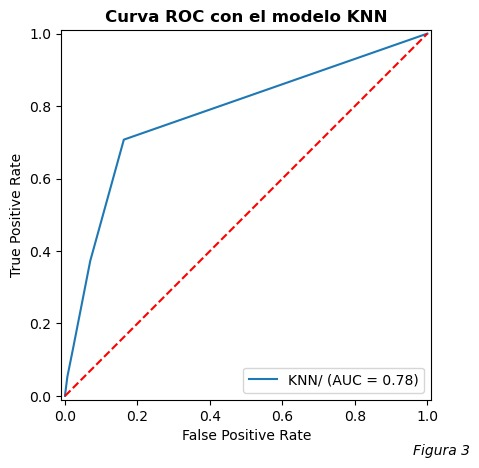
Los resultados indican que el salario predicho aumenta con la edad hasta un cierto punto y luego comienza a disminuir, reflejando una relación no lineal. Además, la educación y la cantidad de horas trabajadas tienen un efecto positivo y significativo sobre el salario, mientras que ser mujer o participar en el PET se asocia con menores ingresos.

4) Gráfico de dispersión



El gráfico muestra que el modelo 5 predice que los salarios tienden a aumentar con la edad hasta un punto máximo (~40-50 años) y luego disminuyen gradualmente, hasta los ~65 que hay una importante baja. Sin embargo, la alta dispersión y la presencia de valores negativos indican que el modelo podría beneficiarse de restricciones adicionales para hacer predicciones más realistas.

1. *Métodos de Clasificación y Performance  
   *

**

La comparación entre las curvas ROC de los modelos Logit y KNN revela un desempeño similar, aunque con diferencias clave. El modelo Logit presenta un AUC de 0.80 y una curva suave, lo que indica una buena capacidad de discriminación entre clases y una predicción probabilística bien calibrada. En contraste, el modelo KNN muestra un AUC ligeramente inferior (0.78) y una curva escalonada, típica de modelos basados en votación discreta como KNN. Aunque ambos modelos tienen un rendimiento aceptable, el Logit ofrece mayor estabilidad y precisión en la estimación de probabilidades, mientras que KNN puede ser más sensible al ruido y depende del valor de *k*. En conjunto, el modelo Logit se perfila como la opción ligeramente superior en términos de capacidad predictiva y consistencia.

Como el accuracy del modelo Logit es de 0.948, es decir, mayor al de KNN (0.934), utilizamos el primero para calcular los valores del punto 6). Aunque también es cierto que, el modelo Logit no predice ningún valor =1 por lo que no es ideal. El de KNN predice valores iguales a 1, sean estos falsos o verdaderos.

Por eso vemos que en el punto 6:

La cantidad de desocupados predecidos son: 0

La cantidad de desocupados encuestados son: 0

La proporción de las personas que no respondieron, pero son y se predijeron como desocupadas es del 100%.

Mientras que si utilizamos el modelo KNN:

La cantidad de desocupados predecidos son: 7 (falso positivo)

La cantidad de desocupados encuestados son: 0

La proporción de las personas que no respondieron pero son y se predijeron como desocupadas es de 0%